The RWTH Large Vocabulary Arabic
Handwriting Recognition System
OpenHaRT 2013 Workshop, Washington DC

Mahdi Hamdani¹, **Patrick Doetsch**¹, Michal Kozielski¹,
Hendrick Pesch¹, Amr El-Desoky Mousa¹
Hermann Ney¹,²

¹Lehrstuhl für Informatik 6
Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University
D-52056 Aachen, Germany

²Spoken Language Processing Group
LIMSI CNRS, Paris, France

Aug. 23, 2013
Outline

- Introduction
  - State of the Art
  - Feature Extraction
  - Visual Model
  - Language Model
- Results
- Conclusions and Future Work
Introduction

What are the used databases?

OpenHaRT database

- Large Arabic Handwriting database
- Pages of handwritten text
- Pages are segmented into words and lines
- Paragraphs are typically multiple lines
Recognition

Image Input

Feature Extraction

Global Search:
maximize
Pr(w_1...w_N) \cdot Pr(x_1...x_T | w_1...w_N)
over w_1...w_N

Recognized Word Sequence

Character Inventory

Writing Variants Lexicon

Language Model
Arabic specificities

- Right to left cursive writing: 28 base characters
- Ligatures, diacritics optional in handwriting!
- Letter can have many shapes (position dependent)

(a) Ligatures
(b) Diacritics
State of the art

Arabic Handwriting Recognition Competition

- ICDAR (2005, 2007, 2009 and 2011) and ICFHR 2010
  - Best systems are based on Hidden Markov Models (HMMs) or Long Short Term Memory (LSTM) Neural Networks
  - Graves and Schmidhuber, Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks, NIPS 2008 (TU Munich)
  - Doetsch et al. Comparison of Bernoulli and Gaussian HMMs using a vertical repositioning technique for off-line handwriting recognition, ICFHR 2012 (RWTH Aachen)

NIST 2010 OpenHaRT evaluation

- Best system based on HMMs
Feature Extraction

Appearance-Based

- Images are scaled to the same height
- Recognition of characters within a context
- Sliding window, PCA reduction
- Typically: large context-window with maximum overlap
Feature Extraction...

Window Repositioning Technique

- Images Scaling
- Center of gravity calculation
- Window repositioning
- Features are gray scale pixel values
Visual Model
Context Dependent Model

Cart Decision tree

▶ Nodes are tagged with questions
▶ Leaves are tagged with class labels
▶ Questions concern the visual classes

<table>
<thead>
<tr>
<th>Types</th>
<th>Examples of Characters</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Ascenders</td>
<td>ن،ش،س</td>
<td>نُسُس</td>
</tr>
<tr>
<td>Descenders</td>
<td>ز،ر،و</td>
<td>زرو</td>
</tr>
<tr>
<td>Occlusions</td>
<td>ف،ة،ه</td>
<td>فةه</td>
</tr>
</tbody>
</table>
Discriminative training

Motivation

HMM Training Approaches

- Maximum Likelihood (ML) training
  - Separate model constructed for each class
  - Only in-class information is available!

- Discriminative training
  - Aim is to separate the classes
  - Classifier performance reflected by objective function
  - Training Criterion: e.g. Minimum Phone Error (MPE)

Framewise Training Approaches

- (Discriminative) training with fixed segmentation
- Segmentation has to be provided
- Here: Neural networks
Discriminative HMM training

Objective Function

Definitions

- \( X \): Sequence of observation vectors over time
- \( W \): Written word sequence
- \( p_\Lambda(W|X) \): Class posterior distribution with parameter set \( \Lambda \)

Training

- Training examples \((X_r, W_r)\)
- Criterion

\[ \hat{\Lambda} = \arg \min_\Lambda \left\{ \sum_{r=1}^{R} L[p_\Lambda(W_r|X_r)] + \text{reg\_term}(\Lambda) \right\} \]

with a loss function \( L[p_\Lambda(W_r|X_r)] \)
Discriminative training

Training Criterion...

Minimum Phone Error (MPE)

\[ L^{(\text{MPE})}[p_{\Lambda}(X_r, \cdot), W_r] = \sum_{W \in \cdot} E(W, W_r) \frac{p_{\Lambda}(X_r, W)^\gamma}{\sum_{V} p_{\Lambda}(X_r, V)^\gamma} \]  

(1)

\( E(W, W_r) \): Measure of correctly transcribed characters in \( W \)

\( X_r \): observation vector sequence

\( W_r \): transcription word sequence

\( \gamma \): approximation level (controls the smoothness of the criterion)
Framewise training with recurrent neural networks
Non-linear feature transformation
Full context modeling through bidirectional topology
Long Short Term Memory (LSTM)

- Constant error flow without loss of short time lag capability
- Replace units by memory cells

- Input Gate: Protects error flow inside cell from irrelevant inputs
- Output Gate: Protects error flow of other cells from irrelevant inputs
- Forget Gate: Provides a way to reset cell state
ANN/HMM Combination

▶ Idea: Train neural network on aligned feature vectors $(x^T_1, s^T_1)$
  ▶ Alignment must be provided by previously trained HMM
▶ Tandem: Use posterior probabilities as features to train HMM
▶ Posterials highly correlated: LDA/PCA to $n$ dimensions
  ⇒ Requires full retraining of HMM
▶ Hybrid: Use posterior probabilities as state emission probability of HMM

$$p_t(x_t|s) = \frac{p_t(s|x_t)}{p(s)^\alpha}$$

▶ $\alpha$: Priori scaling factor
MADA toolkit

- Morphological Analysis and Disambiguation for Arabic (Habash et al. 2009)
- Tool for morphological and contextual analysis of raw Arabic text
- Examines all possible analyses for each word
- Selects the analysis that matches the current context best
- Tokenize the disambiguated text generated by MADA
- A tokenization scheme is provided
Language Model Training...

MADA Toolkit

Different variables for the tokenization scheme

<table>
<thead>
<tr>
<th>Type</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefix</td>
<td>QUES</td>
<td>The “question“ proclitic (e.g. آ)</td>
</tr>
<tr>
<td></td>
<td>CONJ</td>
<td>”Conjunction“ proclitic (و and ٰ)</td>
</tr>
<tr>
<td></td>
<td>PART</td>
<td>“Article“ proclitic</td>
</tr>
<tr>
<td></td>
<td>FUT</td>
<td>Future marker (س)</td>
</tr>
<tr>
<td></td>
<td>NART</td>
<td>Negative articles only (لا and مَا)</td>
</tr>
<tr>
<td></td>
<td>DART</td>
<td>Definite article (ال)</td>
</tr>
<tr>
<td>Radical</td>
<td>REST</td>
<td>Remainder of the word</td>
</tr>
<tr>
<td>Suffix</td>
<td>PRON</td>
<td>Enclitics</td>
</tr>
</tbody>
</table>

For example, the word أو وسيَكَابَتَهَا will be decomposed to و (conjunction) + ُـهَا (future marker clitic) + يكَابَتـُهَا (rest) + يُـلاَم (suffix).
Vocabulary Selection

- The $M$ most frequent full-words are not decomposed
- The selected vocabulary contains new elements which are prefixes, suffixes and stems
- Prefixes and suffixes are tagged with a special marker ("" + "")
- Recognition of unknown words is possible by combining the vocabulary elements

Language Model

- Collected in domain text
- Decomposition using MADA toolkit
- Standard $n$-gram LM trained using the SRILM toolkit (Stolcke et al. 2012)
Results

Tasks

Arabic Handwriting Recognition Competition

- IfN/ENIT database
- Limited vocabulary size

OpenHaRT 2013 evaluation

- Constrained task
  - LM restricted to the training text
- Unconstrained task
  - Additional data used for the LM training
  - The used data is collected from publicly available newspapers and web-forums
  - 1 billion running words
## Results on the IfN/ENIT task

<table>
<thead>
<tr>
<th>System</th>
<th>WER [%]</th>
<th>CER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHMM, MLP Tandem (ICDAR’11)</td>
<td>5.9</td>
<td>4.7</td>
</tr>
<tr>
<td>GHMM, MLP Hybrid (ICDAR’11)</td>
<td>10.3</td>
<td>8.1</td>
</tr>
<tr>
<td>GHMM</td>
<td>13.1</td>
<td>10.6</td>
</tr>
<tr>
<td>+ Repo.</td>
<td>6.4</td>
<td>4.6</td>
</tr>
<tr>
<td>GHMM, LSTM Tandem</td>
<td>7.2</td>
<td>5.6</td>
</tr>
<tr>
<td>+ Repo., [Doetsch et al., 2012]</td>
<td>4.8</td>
<td>3.7</td>
</tr>
<tr>
<td>BHMM, UPV, [Doetsch et al., 2012]</td>
<td>6.2</td>
<td>-</td>
</tr>
<tr>
<td>MD-LSTM, TUM, [Graves et al., 2009]</td>
<td>6.6</td>
<td>-</td>
</tr>
</tbody>
</table>
## Results...

### OpenHaRT Data Description

<table>
<thead>
<tr>
<th>Table: OpenHaRT data statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td># of pages</td>
</tr>
<tr>
<td># of paragraphs</td>
</tr>
<tr>
<td># of words</td>
</tr>
<tr>
<td># of characters</td>
</tr>
<tr>
<td>avg number words/paragraph</td>
</tr>
<tr>
<td>avg number characters/word</td>
</tr>
</tbody>
</table>
Results on OpenHaRT for the constrained task

Table: Results of the RWTH handwriting recognition system on the OpenHaRT constrained task

<table>
<thead>
<tr>
<th>System</th>
<th>Vocabulary size</th>
<th>WER [%]</th>
<th>CER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>99k</td>
<td>27.4</td>
<td>10.9</td>
</tr>
<tr>
<td>Sub-lexical approach</td>
<td>94k</td>
<td>26.8</td>
<td>10.1</td>
</tr>
</tbody>
</table>

OOV rates

- Baseline full-words: 8.29%
- Sub-lexical: 5.70%
Results...

Results on OpenHaRT for the unconstrained task

- Vocabulary size: 200k
- OOV: 3.5%

Table: Results of the RWTH handwriting recognition system on the OpenHaRT unconstrained task

<table>
<thead>
<tr>
<th>System</th>
<th>WER [%]</th>
<th>CER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHMM CI</td>
<td>33.2</td>
<td>15.4</td>
</tr>
<tr>
<td>GHMM CD</td>
<td>25.9</td>
<td>10.1</td>
</tr>
<tr>
<td>+BLSTM</td>
<td>19.9</td>
<td>5.9</td>
</tr>
<tr>
<td>+MPE</td>
<td>17.0</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

Conclusions

▶ Morphological Decomposition using MADA toolkit
▶ Improvement up to 1% in the constrained task
▶ Same results of the baseline system in the unconstrained task
▶ Lexicon flexibility with competitive results in very large vocabulary task

Future Work

▶ Comparison of the morphological decomposition with other types of decomposition (e.g. part of words)
▶ Combination with character based language models
Thank you for your attention

Mahdi Hamdani, Patrick Doetsch, Michal Kozielski, Hendrick Pesch, Amr El-Desoky Mousa Hermann Ney

hamdani@i6.informatik.rwth-aachen.de